Bayesian Data Analysis

Christophe

December 7, 2015

2.1

Suppose we have a four-sided die from a board game. On a tetrahedral die, each face is an equilateral triangle. When you roll the die, it lands with one face down and the other three faces ivsible as three-sided pyramid. The faces are numbered 1-4, with the value of the bottom face printed (as clustered dots) at the bottom edges of all three visible faces. Denote the value of the bottom face as x. Consider the following three mathematical descriptions of the probabilities of x. Model A: $p(x) = 1/4$. Model B: $p(x) = x/10$. Model C: $p(x) = 12/(25x)$. For each model, determine the value of $p(x)$ for each value of x. Describe in words what kind of bias (or lack of bias) is expressed by each model.

```
model_a <- NULL
model_b <- NULL
model_c <- NULL

for (i in 1:4){
  x <- sides <- seq(1:4)
  model_a[i] <- (1/4)
  model_b[i] <- (x[i]/10)
  model_c[i] <- (12/(25*x[i]))
}

model_table <- rbind(model_a,model_b,model_c)
colnames(model_table) <- lapply(sides, function(pre,post) paste(pre,post), pre="Side" )
kable(model_table)
```

	Side 1	Side 2	Side 3	Side 4
model_a	0.25	0.25	0.25	0.25
model_b	0.10	0.20	0.30	0.40
model_c	0.48	0.24	0.16	0.12

> As we can see all elements of `model_a` are the same so all outcomes have equal probability. However, in `model_b` and `model_c` not all of the elements have equal probability and the ones with the highest probability have a bias to be selected more often the the other sides.

5.1

This exercise extends the ideas of Table 5.4, so at this time, please review Table 5.4 and its discussion in the text. Suppose that the same randomly selected person as in Table 5.4 gets re-tested after the first test results was positive, and on the re-test, the result is negative. When taking into account the results of both test, what is the probability that the person has the diease?

```
x <- positivetest_hasdisease <- .99
y <- positivetest_nodisease <- .05
z <- general_disease <- .001

positivetest_withdiease <- (x * z / (x * z + (y * (1 - z))))
percent(positivetest_withdiease)
```

[1] "1.94%"

```
have_disease <- positivetest_withdiease

positive_disease_withnegativetest <- (( 1 - x) * have_disease/ ((1 - x) *
                                        have_disease + (1 - y) *
                                          (1 - have_disease)))

percent(positive_disease_withnegativetest)
```

[1] "0.0209%"

5.2

(A) Suppose that the population consists of 100,000 people. Compute how many people would be expected to fall into each cell of Table 5.4. To computer the expected frequency of people in a cell, just multiply the cell probability by the size of the population. To get you started, a few of the cells of the frequency table are filled in here:

	$\theta =$ ☹	$\theta =$ ☺	
D = +	$\text{freq}(D{=}+, \theta{=}$☹$)$ $= p(D{=}+, \theta{=}$☹$)\,N$ $= p(D{=}+ \mid \theta{=}$☹$)\,p(\theta{=}$☹$)\,N$ $= 99$	$\text{freq}(D{=}+, \theta{=}$☺$)$ $= p(D{=}+, \theta{=}$☺$)\,N$ $= p(D{=}+ \mid \theta{=}$☺$)\,p(\theta{=}$☺$)\,N$ $=$	$\text{freq}(D{=}+)$ $= p(D{=}+)\,N$ $=$
D = -	$\text{freq}(D{=}-, \theta{=}$☹$)$ $= p(D{=}-, \theta{=}$☹$)\,N$ $= p(D{=}- \mid \theta{=}$☹$)\,p(\theta{=}$☹$)\,N$ $= 1$	$\text{freq}(D{=}-, \theta{=}$☺$)$ $= p(D{=}-, \theta{=}$☺$)\,N$ $= p(D{=}- \mid \theta{=}$☺$)\,p(\theta{=}$☺$)\,N$ $=$	$\text{freq}(D{=}-)$ $= p(D = -)\,N$ $=$
	$\text{freq}(\theta =$ ☹$)$ $= p(\theta =$ ☹$)\,N$ $= 100$	$\text{freq}(\theta =$ ☺$)$ $= p(\theta =$ ☺$)\,N$ $= 99,900$	N $= 100,000$

Notice the frequencies on the lower margin of the table. They indicate that out of 100,000 people, only 100 have the disease, while 99,900 do not have the disease. These marginal frequencies instantiate the prior probability that $p(\theta =: () = 0.001$. notice also the cell frequencies in the column $\theta =:)$ which indicate that of 100 people with the dease , 99 have a positive test result and 1 has negative test result. These cell frequencies instantiate the hit rate of 0.99. Your job for this part of the exercise is to fill in frequencies of the remaining cells of the table.

```
x <- positivetest_hasdisease <- .99
y <- positivetest_nodisease <- .05
z <- general_disease <- .001
n <- 100000

Dtheta <- list()
Dtheta[1] <- (x * z) * n
Dtheta[2] <- (1 - x) * z * n
Dtheta[3] <-  Dtheta[[1]] + Dtheta[[2]]

NDtheta <-  list()
NDtheta[1] <- (y * (1-z)) * n
NDtheta[2] <- (1 - y) * (1 - z) * n
NDtheta[3] <- NDtheta[[1]] + NDtheta[[2]]

total <- list()
for (i in 1:3) {
total[i] <- Dtheta[[i]] + NDtheta[[i]]
}

a <- list.rbind(Dtheta)
b <- list.rbind(NDtheta)
c <- list.rbind(total)
df <- data.frame(a,b,c)
colnames(df) <- c("Has Disease", "No Disease", "Total")
rownames(df) <- c("Test Positive", "Test Negative", "Total")
kable(df)
```

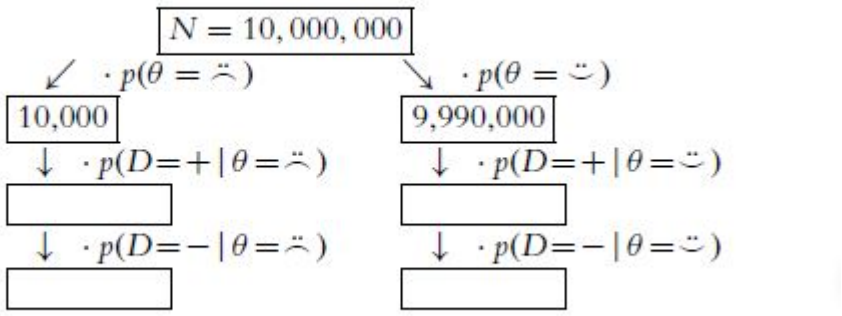
	Has Disease	No Disease	Total
Test Positive	99	4995	5094
Test Negative	1	94905	94906
Total	100	99900	100000

(B) Take a good look at the frequencies in the table you just computed for the previous part. These are the so-called "natural frequencies" of the events, as opposed to the somewhat unintuitive expression in terms of conditional probabilities (Gigerenzer & Hoffrage, 1995). From the cell frequencies alone, determine the proportion of people who have the disease, given that their test result is positive. Before computing the exact answer arithmetically, first give a rough intuitive answer merely by looking at the relative frequencies in the row $D = +$. Does your inuitive answer match the intuitive answer you provided when originally reading about Table 5.4? Probably not. Your intuitive answer here is probably much closer to the correct answer. Now compute the exact answer arithmetically. It should match the result from applying Bayes' rule in Table 5.4.

> The `D = +` is `Positive Test` and 99 of 5,094. A rough answer is `100` divided by `5000` which is approximately 2%. It does match my answer, I remembered a cancer problem from earlier in class where the rates of cancer from a positive test where much lower than expected. If I solve this question for the exact answer it is 1.94%.

(C) now we'll consider a related representation of the probabilities in terms of natural frequencies, which is especially useful when we accumlate more data. This type of representation is called a "Markov" representation by Krauss, Martignon, and Hoffrage (1999). Suppose now we start with a population of N = 10,000,000 people. We expect 99.9% of them (i.e. 9,990,000) not to have the disease, and just 0.1% (i.e. 9,900) will be expected to test positive. OF the 9,990,000 people who do not have the diease, 5% (i.e., 499,500) will be expected to test positive. Now consider re-testing everyone who has tested positive on the first test. How many of them are expected to show a negative result on the retest?

Use the diagram to compute your answer:



When computing the frequences for the empty boxes above, be careful to use the proper conditional probabilities!

```
x <- positivetest_hasdisease <- .99
y <- positivetest_nodisease <- .05
n <- 10000000

LB1 <- ((n/1000) * x)
LB2 <- LB1 * (1-x)
Left_Branch <- list(LB1,LB2)

RB1 <- (n-10000) * y
RB2 <- RB1 * (1 -y)
Right_Branch <- list(RB1, RB2)

df <- as.matrix(Left_Branch)
df <- cbind(df, Right_Branch)
colnames(df) <- c("Left Branch", "Right Branch")

kable(df)
```

Left Branch	Right Branch
9900	499500
99	474525

(D) Use the diagram in the previous part to answer this: What proportion of people who test positive at first and then negative on restest, actually have the disease? In other words, of the total number of people at the bottom of the diagram in the previous part (those are the people who tested positive then negative), what proportions of them are in the left branch of the tree? How does the result compare with your answer to Exercise 5.1?

```
paste("The proportion of those that tested positive first and then negative on retest are ",
      percent(LB2/(LB2+RB2)))
```

[1] "The proportion of those that tested positive first and then negative on retest are 0.0209%"